Precision of genetic relationship estimates based on molecular markers

José F. Barbosa-Neto, Carlos M. Hernández, Louise S. O'Donoughue, Mark E. Sorrells*

J.B. Barbosa-Neto and M.S. Sorrells, Dep. of Plant Breeding and Biometry, 252 Emerson Hall, Cornell University, Ithaca, NY 14853-1902; C.M. Hernandez, Biometrics Unit, 324 Warren Hall, Cornell University, Ithaca, NY 14853-1902; L.S O'Donoughue, Agric. Canada Res. Branch, Plant Res. Center, Central Exp. Farm, Bldg. 50, Ottawa, Ontario, K1A OC6, Canada. Paper number 829 of the Plant Breeding Series.
*Corresponding author (mark_sorrells@qmrelay.mail.cornell.edu).

BU-1290-M

May, 1995

ABSTRACT

Genetic progress through selection is directly related to the amount of variability present in the population and the quality of genes contributed by the parents. Molecular markers can be used for estimating genetic relationship between potential parents. A statistical methodology using the size of a $(1-\alpha)\%$ confidence interval was developed to determine the precision in the estimation of genetic distance between pairs of cultivars. Precision of relationship estimates was affected by type of genetic index used, number of cultivars, and amount of genetic diversity present in the studied group. The size of the $(1-\alpha)\%$ confidence interval decreased as the number of RFLP fragments increased. Oat and wheat diversity studies were used to illustrate the methodology.

INTRODUCTION

Cultivar development is based on the selection of superior individuals in segregating populations. Genetic progress through selection is directly related to the amount of variability present in the population and the quality of genes contributed by the parents; as a consequence, the correct choice of parents can maximize the genetic variability and the average performance of a segregating population. Breeders choose parents mainly based on pedigree information and mean performance.

Genetic relationship between cultivars can be used for selection of parents in a breeding program (Frei *et al.*, 1986). Quantitative and qualitative traits have been used to assess genetic diversity between cultivars in oat (Souza and Sorrells, 1991a; Souza and Sorrells, 1991b) and maize (Smith *et al.*, 1990). Molecular markers have been used to predict genetic diversity and combining ability between potential parents. Smith *et al.* (1990) estimated similarities among a group of elite maize (*Zea mays* L.) inbreds and concluded that restriction fragment length polymorphism (RFLP) information could be used to predict combining ability. On the other hand, Melchinger *et al.* (1990) did not detect a correlation between RFLP based distance and combining ability, although the authors recommended RFLP to assign cultivars to different genetic pools. Martin *et al.* (1995) studied the relationship between hybrid performance and parental diversity measured using polymerase chain reaction (PCR) analysis in wheat (*Triticum aestivum* L. em Thell.). They compared seven hard red spring wheats and concluded that PCR analysis using 27 polymorphic primers was not useful for heterosis prediction. Studies on genetic diversity between cultivars using molecular markers have been conducted in several other crops, such as barley (*Hordeum vulgare* L.)

(Melchinger *et al.*, 1994), rice (*Oryza sativa* L.) (Wang *et al.*, 1992), and soybean [*Glycine max* (L.) Merr.] (Keim *et al.*, 1992).

Advantages of DNA marker based genetic relationship over other measures include high polymorphism and detection of allelic variation by descent and in state. The disadvantages are the cost and time required for an RFLP study. A practical method is needed to determine the number of RFLP required to achieve a given level of precision for relationship between two cultivars. Typically, 25 (Wang *et al.*, 1992) to 640 (Keim *et al.*, 1992) probe-restriction enzyme combinations have been used. Keim *et al.* (1992), working with soybean, used the standard error of the distance estimation between a pair of cultivars as a criteria to determine the number of probes required. Tivang *et al.* (1994) used a coefficient of variation associated with a genetic distance estimation as a measure of precision. However, their calculations were not generalized for other distance indexes and plant species.

The objective of this study was to develop a statistical approach to determine the number of probe-restriction enzyme combinations necessary to estimate genetic distance between cultivars with a given confidence interval. The methodology was applied to a sample of 34 wheat cultivars using RFLP and coefficient of parentage to assess genetic relationship.

THEORY

Measures of genetic distance

Several indices have been proposed to measure genetic distance. Some indices, such as Nei's distance, include explicit statements about underlying genetic models. On the other hand, geometric distances do not involve any genetic concept (Weir, 1990). They are based on Euclidean distance as follows:

$$d(\text{cultivar 1, cultivar 2}) = [\Sigma_i^N (p_{i1}-p_{i2})^2]^{1/2},$$

where p_{i1} = allelic frequency of the i^{th} RFLP fragment in cultivar 1, p_{i2} = allelic frequency of the i^{th} RFLP fragment in cultivar 2, and N is the number of RFLP fragments scored in both cultivars.

In this study, a Euclidean distance index was used in order to differentiate cultivars; because phylogenetic relationship is not relevant for cultivars and there was no need for genetic concepts, such as mutation rate or time of divergence. The distance index (DI) used was:

$$DI = [\Sigma_i^N (p_{i1}-p_{i2})^2/N]^{1/2}.$$

This index ranges from zero, when cultivars are identical at probed locations, to one, for completely dissimilar cultivars at these locations. It is similar to Modified Roger's distance (Wright, 1978); however, RFLP fragments instead of loci were used due to the phenomenon of multiple dose of fragments in polyploid species. Multiple fragments hinder determination of allelic associations because a single probe may hybridize to loci on more than one chromosome (Sorrells, 1992).

Statistical distribution of distance index

The real frequency of mismatches (ϕ) between two cultivars can be determined only at the level of nucleotide sequencing, where total DNA is sequenced for both cultivars. RFLP fragments and DI are used to estimate ϕ. For inbred cultivars, the allelic frequency of RFLP fragments for any cultivar can be either zero (absence) or one (presence). As a consequence, the value of the expression $(p_{i1}-p_{i2})^2$ will be either zero or one and assuming independence between RFLP fragments the summation of $(p_{i1}-p_{i2})^2$ over RFLP fragments is distributed as a Binomial (N, ϕ). The estimator DI is the sample mean of this binomial distribution which has mean ϕ and variance $\phi(1-\phi)/N$. In addition, by the Central Limit Theorem DI is

asymptotically normally distributed with mean ϕ and variance $\phi(1-\phi)/N$. An important assumption that was made to establish the statistical distribution is that each RFLP fragment were independent and identically distributed. This is not always true, since an RFLP probe that hybridizes to a fragment that contains an internal restriction site for the restriction enzyme used could result in two visible fragments.

Sample size required to estimate distance between cultivars

In order to determine how many probes have to be surveyed to estimate distance between two cultivars, it is necessary to establish the degree of precision required for the genetic distance index (DI). One approach is to use the size of a $(1-\alpha)\%$ confidence interval (CI) for DI, where $(1-\alpha)$ is the probability that the CI covers the real value of DI. Since DI is a square root of the observed frequency of mismatches (f) and this transformation function is monotone in the interval [0, 1]; it is possible to estimate a confidence interval for ϕ and transform it for DI (Mood et al., 1974). Let the size of a $(1-\alpha)\%$ CI for ϕ be 2E. Since

$$E = (Z_{\alpha/2}\sigma)/N^{1/2},$$

where E = half size of a $(1-\alpha)\%$ CI for ϕ, $Z_{\alpha/2}$ = value of the standard Normal with $\alpha/2$ probability, σ = standard deviation of f, estimator of ϕ, and N = total number of RFLP fragments scored; it follows that after substituting σ for the asymptotic standard deviation:

$$E = Z_{\alpha/2}[f(1-f)/N]^{1/2},$$

where f is the observed frequency of mismatches. Solving for N in the above expression:

$$N = [(Z_{\alpha/2})^2 f(1-f)]/(E^2),$$

it is possible to calculate the number of RFLP fragments that need to be scored in order to estimate the genetic distance between two cultivars,

given the desired accuracy (E), the level of certainty (α), and the number of RFLP fragment mismatches observed. Table 1 contains the conversion values of E for f that correspond to E = 0.10, 0.05, and 0.01 for DI.

As an example, assume a comparison between two inbreds. Based on pedigree, previous research, or a preliminary survey, f is set as 0.25. The experiment requires a 90% CI for DI with a half size E = 0.10. Using Table 1 for DI = 0.50 , an estimate for the number of scored fragments required is:

$N = [1.64^2 \times 0.25 \times (1 - 0.25)] / 0.0980^2$

$N = 52.51$

Thus, for two cultivars, approximately 53 RFLP fragments must to be scored to achieve the desired precision.

MATERIAL AND METHODS

Germplasm

Thirty-four wheat cultivars were surveyed for pedigree and RFLP. This group consisted of winter and spring types and both public and proprietary inbreds. In addition to the wheat cultivars, RFLP surveys from 84 oat (*Avena sativa* L.) cultivars were used in order to exemplify the statistical methodology proposed. Methodology and results for the oat study were reported by O'Donoughue et al. (1994).

Coefficient of parentage

Pedigrees were obtained from Zeven and Zeven-Hissink (1976) and Matthews and Anderson (1995). The pedigree for a cultivar was traced back to landraces or to parental lines with unknown parentage. Coefficients of parentage (COP) were estimated as described by Cox et al. (1985). The following assumptions were made: 1) each parent contributed equally to the offspring and 2) there was no selection during inbreeding.

The relationship between a cultivar and its parents was assumed to be r_p = 0.5 and a re-selection of a previous cultivar or landrace had an assigned relationship of r_p = 0.75. Cultivars with unknown origin were assumed to be unrelated (r_p = 0) and the relationship of a cultivar with itself was r_p = 1.

RFLP analysis

A bulk sample of leaf tissue was harvested for DNA extraction from at least 10 greenhouse grown plants that were 3-4 weeks old. The tissue was chilled in liquid nitrogen and ground to fine powder. DNA extraction, DNA digestion, and Southern blotting were done according to Anderson et al. (1992). The restriction enzyme Eco-RI was used for DNA digestion. Hybridization procedures were described by Heun et al. (1991), except exposures of autoradiograms were 2-12 days.

Thirty-four probes were selected from the wheat arm map published by Anderson et al. (1992) to represent loci distributed over the entire genome. A barley cDNA library (22 probes), an oat cDNA library (9 probes), and a wheat genomic library (3 probes) were used on the 34 wheat cultivars. All fragments hybridizing to a probe were scored presence (1) or absence (0). If all fragments from a probe were monomorphic for the set of cultivars they were excluded from the study.

Statistical analysis

Genetic distances for each pair of cultivars were calculated from a matrix constructed with the RFLP data, where rows contained the genotypes and columns, the fragments scored. Distance index estimations were calculated using NTSYS-pc (Rohlf, 1992). The observed frequency of mismatches (f) for both the wheat and oat studies were estimated by averaging the f obtained for each pair of cultivars. In addition to the total

group, subgroups with 10 or 20 randomly chosen cultivars were used to provide examples for the proposed methodology.

In the wheat cultivars study, cluster analysis was performed for both COP and DI. Dendograms were constructed based on the average linkage process. DI and COP matrixes were correlated and compared using the normalized Mantel statistic, which assumes bivariate normal distribution, but not independence between pairs of observations.

RESULTS AND DISCUSSION

Methodology for RFLP sample size estimation

Confidence intervals were calculated for a range of DI and population sizes (Table 2). For illustration, 90% CI was used because other $(1-\alpha)\%$ CI can be estimated by changing the $Z_{\alpha/2}$ value. The number of fragments that need to be scored in an RFLP study increases rapidly as the CI size (2E) decreases. The use of an E = 0.01 requires data for a large number of fragments, even when the number of comparisons is very small. On the other hand, the number of fragments required for an E = 0.10 are feasible for the entire range of DI. The CI size where E = 0.05 has an intermediate position, requiring reasonable sample sizes when comparing two or three cultivars or when the estimated value of DI is over 0.60.

The parameter ϕ reflects the genetic diversity of the population under study. Populations composed of related inbred lines generally have larger DI variances compared to populations of distantly related lines. This is due to the square root transformation on the binomial distribution, which makes the variance of DI inversely proportional to the frequency of mismatches observed (f). Fig. 1 depicts the effect of different ϕ on the size of a 90% CI for DI. When comparing lines that are related, many RFLP fragments are monomorphic but results in a larger variance estimate. As a

consequence, probes with high polymorphism information content (PIC) are preferred, because the frequency of mismatches will be higher, resulting in a lower variance. This variance reflects the estimation error that would be detected in the case of repetitive sampling with different RFLP probes.

The $(1-\alpha)\%$ CI estimation is used to assess the degree of precision attained in the estimation of the real DI between cultivars. When more than two cultivar are studied, it is important to decide if pairwise or simultaneous comparison will be used. Pairwise comparison is generally used in studies that attempt to correlate genetic distances with other measures, such as coefficient of parentage or heterosis. The correlation coefficient precision is affected by the number of pairs that are being correlated and not by the precision of the genetic distance estimator. On the other hand, when a phylogenetic tree is constructed, precise estimation of genetic distance between all pair of cultivars studied is a primary requirement. For simultaneous comparison the number of cultivars affects the number of fragments necessary for accurately estimating genetic relationship (Table 2). To extend the results to more than two cultivars one must consider $M = [L(L-1)/2]$ different pairwise comparisons between L inbreds. An optimal number of fragments to achieve a desired precision E will affect the probability of not achieving this precision for all M comparisons. Therefore, the α becomes:

$$1-(1-\alpha')^{M}.$$

The pairwise α'-level must be such that the expression above is equal to the desired overall α-level. For example, when comparing line A against lines B and C, two comparison are made. If an α'-level of 0.05 is used in each comparison, the overall α-level will be approximately 0.10.

Estimation of ϕ

For a set of lines not previously studied, the parameter ϕ is unknown for a population and has to be estimated in order to calculate the number of RFLP fragments that need to be surveyed. There are at least three different ways to obtain an estimate for ϕ. The coefficient of parentage between two inbreds could be used as a preliminary estimate or one might obtain a preliminary estimate from previous studies. A more refined approach would be to survey an initial set of probes on the population under study and to use the observed number of mismatches (f) as an estimate for ϕ. This preliminary value could be used to estimate the required sample size and then updated as soon as more information is available for the population.

When comparing two cultivars there is only one possible value for f; however, in a study involving more than two cultivars there is one value for each pair being compared. The use of an f value averaged over all comparisons for the formula to obtain N seems to be a logical approach. On the other hand, if the researcher wants to be more or less stringent, it is possible to use the minimum or maximum observed value of f, respectively.

Generalization to other distance measures

A generalization of this methodology to other distance measures can be done using the Central Limit Theorem to approximate the distribution for the index and a $(1-\alpha)\%$ CI can be estimated. For of inbred lines, where the allelic frequency is always considered to be one or zero, distance or similarity indexes are based on Bernoulli distribution. On the other hand, when the study is conducted with varieties that are not homozygous, the

Bernoulli distribution may not apply and the basic statistical distribution has to be derived.

Depending on the dissimilarity within a population, the number of RFLP fragments that need to be scored to give the same (1-α)% CI for different distance measures may also be different. For example, the genetic distance indexes DI and Nei's are non-linear transformations of the frequency of mismatches parameter, ϕ (Fig. 2). In a population with high dissimilarity between lines, Nei's distance value tends to be infinite, which would require a large number of RFLP fragments to achieve a reasonable confidence interval. On the other hand, the distance index used in this study, DI, requires larger sample sizes for populations with higher similarity. Roger's distance and Modified Roger's distance exhibit the same behavior as DI, because these indexes are based on the same Euclidean distance but are weighted by the number of loci scored rather than the number of RFLP fragments.

Uses of the proposed methodology

The proposed methodology allows a researcher to chose an optimal number of probes to achieve a given level of precision for the estimated distance index. In order to use the methodology, the average number of mismatches, f, has to be calculated as discussed earlier, and an estimation of the average number of fragments hybridizing to each probe has to be obtained. The number of fragments per probe does not vary considerably from collection to collection within the same species; as a consequence, this value can be obtained from the literature or can be calculated from the preliminary survey on the collection under study. After attaining the preliminary sample size, the updated results are used to recalculate the

sample size required. This process may be terminated when the recalculated sample size is similar to the previous one.

This statistical methodology does not take into account distribution of the probes over the genome. In genetic diversity studies, probes should be chosen to uniformly represent the whole genome, avoiding biases due to sampling. In addition, only one restriction enzyme should be used in each study, since polymorphisms obtained with more than one restriction enzyme may not be independent (Miller and Tanksley, 1990).

Oat study: an example

O'Donoughue et al. (1994) analyzed genetic diversity in a set of 84 oat cultivars using RFLP data. The data set containing scores for 358 RFLP fragments was re-analyzed using DI. A 90% CI and the estimated minimum RFLP fragment sample size to obtain an E = 0.10 are presented in Table 3. Subsets of cultivars were selected at random and the average frequency of mismatches was calculated in order to illustrate the effect of number of comparisons made in the calculation of the CI. The 90% CI ranged from E = 0.07 to E = 0.10 when simultaneous comparison was used. For pairwise comparison the 90% CI was E = 0.04. The number of fragments required for an E = 0.10 in the complete set was 60 and 359 RFLP fragments for pairwise and simultaneous comparison, respectively.

Genetic relationship for a diverse set of wheat cultivars

Table 4 contains characteristics of the 34 wheat cultivars evaluated for genetic relationship in this study. The hard red grain type was the group most represented (79%); while soft red and soft white types represented 15% and 6% of the genotypes sampled, respectively. About half of the cultivars studied were spring types (47%). The number of RFLP fragments detected per probe ranged from three to 10 with an average of

6.13, and the number of different phenotypes per polymorphic clone from two to eight. Only three out of 34 probes (9%) were monomorphic for all cultivars; 68% of the 190 RFLP fragments scored were polymorphic.

The average genetic distance index within the group of 34 wheat cultivars studied was 0.45. A dendogram for this set of 34 wheat cultivars using RFLP data suggested two major groups (Fig. 3). The first group was composed of cultivars B17717 through Tascosa and was more variable than the second one, from Bezostaya through Lar193. The cultivar Arthur was equidistant from both groups. Winter and spring types did not cluster in different groups as observed for barley (Melchinger et al., 1994) and oat (O'Donoughue et al., 1994); the intensive use of winter germplasm in the improvement of spring types in wheat may have contributed to this result. All winter types clustered in the first group were related to Turkey (hard red winter cultivar). In the second group some winter types were related to the landrace Mediterranean. Associations between RFLP clusters and other morphological traits, such as grain color and type and plant height, were not observed. Martin et al. (1995) reported an average similarity index (Nei and Li, 1979) of 0.13 in a group of seven hard red spring wheats. The square root of (1 - 0.87) is equivalent to a DI of 0.36 in their study. The average of 0.45, found in this study, was similar to the oat (DI = 0.40) (O'Donoughue et al., 1994) and barley studies (DI = 0.46) (Melchinger et al., 1994).

Similarly, cluster analysis of the coefficient of parentage (COP) did not separate winter and spring types (Fig. 4). The average COP for the whole group was 0.10, the values ranged from zero to 0.70. It was interesting that COP did not separate winter and spring types to different clusters as observed for oat (Souza and Sorrells, 1991b) and barley (Melchinger et al.,

1994); this result agreed with the RFLP results. Genotypes tended to be grouped according to grain type. Hard red cultivars grouped with at least 12% similarity; however, six hard red genotypes did not share this cluster (Bezostaya, Lac732, Lac519, Lac739, Lar193, and Lar1138). The soft red cultivars Arthur, Caldwell, Pike, and Hart also grouped together with at least 12% similarity. Moro was the cultivar most distantly related by pedigree to any other in the group of 34. This is probably due to the fact that this variety has a large percentage of *Triticum compactum* L. in its parentage.

Comparison of RFLP based genetic distance and coefficient of parentage

The RFLP distance indicated closer genetic relationship between the cultivars than COP. Selection during the process of inbreeding may explain this result, since parents that contribute superior alleles to segregating populations may be over represented in the progeny. O'Donoughue et al. (1994) reported similar results for oat.

The coefficient of correlation between RFLP based genetic distance and COP ($r = -0.35$) was not significantly different from zero according to the normalized Mantel statistic. By definition, COP- and RFLP-based genetic distance are different measures of genetic relationship between two or more genotypes. Coefficient of parentage measures genetic similarity by descent and is subject to several sources of error. Assumptions about equal contribution from each parent, absence of selection during inbreeding, and unrelatedness of ancestor varieties may bias the estimated COP. In addition, unreliable pedigree information or outcrossing for some parental lines may complicate this value. On the other hand, RFLP genetic distance measures the diversity between genotypes by direct sampling and it provides an estimation of alleles alike in state (Graner et al., 1994).

Parents that are used only to improve specific traits, such as disease resistance, may cause a strong bias in the COP estimation, but not in the RFLP-based distance, thus reducing correlation between the two measures. For example, in the study of 34 wheat cultivars, many contained the cross Norin 10/Brevor in the pedigree. This cross is a major source of dwarfing genes in wheat and it has been used specifically with the purpose of reducing stature of new cultivars (Dalrymple, 1980). Autrique (1993) and Graner et al. (1994) also reported low correlation between RFLP based genetic distance and COP in durum wheat (*Triticum durum* L.) and barley, respectively. On the other hand, in maize several authors have reported high correlations between these two measures (Smith *et al.*, 1990; Smith and Smith, 1991; Messmer et al., 1993). Probably, hybrid cultivar development in maize has forced plant breeders to maintain distinct heterotic groups, increasing the diversity between them, which could be detected by either RFLP or COP. In addition, higher frequency of RFLP and better pedigree information in maize may have contributed to the high correlations observed.

Precision of DI estimates in the wheat study

An approach similar to the one used in the oat example was used for the wheat study to evaluate the precision obtained with 190 RFLP fragments scored. Table 5 presents the estimated half size E of a 90% CI for the complete group of cultivars as well for the subsets containing 10 and 20 cultivars. For the complete group of cultivars the desired precision of E = 0.10 was attained with 190 RFLP fragments only for pairwise comparison. In order to attain E = 0.10 for simultaneous comparison it would be necessary to score at least 146 more fragments, or 24 more

probes. When smaller groups of cultivars were compared, the precision of E = 0.10 was obtained.

For precise genetic distance estimates (below E = 0.05), it is necessary to score a large number of RFLP fragments. According to the values of DI estimated in this study, in the oat study (O'Donoughue et al., 1994), and in the barley study (Melchinger et al., 1994), for pairwise comparisons, more than 200 fragments were necessary to be scored to attain an $E < 0.05$. When the required precision of DI estimates increases, as in the case of phylogenetic trees, the required sample size increases to at least 1200 fragments or approximately 196 probes.

REFERENCES

Anderson, J.A., Y. Ogihara, M.E. Sorrells, and S.D. Tanksley. 1992. Development of a chromosomal arm map for wheat based on RFLP markers. Theor. Appl. Genet. 83: 1035-1043.

Autrique, J.E. 1993. Molecular markers applied to wheat improvement. PhD.diss. Cornell University, Ithaca, NY.

Bonierbale, M.W., R.L. Plaisted, and S.D. Tanksley. 1993. A test of the maximum heterozygosity hypothesis using molecular markers in tetraploid potatoes. Theor. Appl. Genet. 86: 481-491.

Cox, T.S., Y.T. Kiang, M.B. Gorman, and D.M. Rodgers. 1985. Relationships between coefficient of parentage and genetic similarity indices in the soybean. Crop Sci. 25: 529-532.

Dalrymple, D.G. 1980. Development and spread of semi-dwarf varieties of wheat and rice in the United States. USDA. Agricultural Economic Report 455.

Frei, O.M., C.W. Stuber, and M.M. Goodman. 1986. Use of allozymes as genetic markers for predicting performance in maize single cross hybrids. Crop Sci. 26: 37-42.

Graner, A., W.F. Ludwig, and A.E. Melchinger. 1994. Relationships among European barley germplasm: II. Comparison of RFLP and pedigree data. Crop Sci. 34: 1199-1205.

Heun, M., A.E. Kennedy, J.A. Anderson, N.L.V. Lapitan, M.E. Sorrells, and S.D. Tanksley. 1991. Construction of a restriction fragment length polymorphism map for barley (*Hordeum vulgare*). Genome 34: 437-447.

Keim, P., W. Beavis, J. Schupp, and R. Freestone. 1992. Evaluation of soybean RFLP marker diversity in adapted germplasm. Theor. Appl. Genet. 83: 205-212.

Martin, J.M., L.E. Talbert, S.P. Lanning, and N.K. Blake. 1995. Hybrid performance in wheat as related to parental diversity. Crop Sci. 35: 104-108.

Matthews, D.E. and O.D. Anderson. 1995. GrainGenes, the Triticeae Genome Database. Internet anonymous ftp. Host probe.nalusda.gov, directory pub/grains.

Melchinger, A.E., A. Graner, S. Mahendra, and M.M. Messmer. 1994. Relationships among European barley germplasm: I. Genetic diversity among winter and spring cultivars revealed by RFLPs. Crop Sci. 34: 1191-1199.

Melchinger, A.E., M. Lee, K.R. Lamkey, and W.L. Woodman. 1990. Genetic diversity for restriction fragment length polymorphisms: relation to estimated genetic effects in maize inbreds. Crop Sci. 30: 1033-1040.

Messmer, M.M., A.E. Melchinger, R.G. Herrmann, and J. Boppenmaier. 1993. Relationships among early European maize inbreds: II. Comparison of pedigree and RFLP data. Crop Sci. 33: 944-950.

Miller, J.C. and S.D. Tanksley. 1990. RFLP analysis of phylogenetic relationships and genetic variation in the genus Lycopersicon. Theor. Appl. Genet. 80: 437-448.

Mood, A.M., F.A. Graybill, and D.C. Boes. 1974. Introduction to the Theory of statistics. 3th ed. McGraw-Hill Publishing Company, New York, NY.

Nei, M. and W. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76: 5256-5273.

O'Donoughue, L.S., E. Souza, S.D. Tanksley, and M.E. Sorrells. 1994. Relationships among North American oat cultivars based on restriction fragment length polymorphisms. Crop Sci. 34: 1251-1258.

Rohlf, F.S. 1992. NTSYS-pc. Numerical taxonomy and multivariate analysis system. Exeter software, Setauket NY.

Smith, J.S.C. and O.S. Smith. 1991. Restriction fragment polymorphisms can differentiate among U.S. mayze hybrids. Crop Sci. 31: 893-899.

Smith, O.S., J.S.C. Smith, S.L. Bowen, R.A. Tenborg, and S.J. Wall. 1990. Similarities among a group of elite maize inbreds as measured by pedigree, F1 grain yield, heterosis, and RFLPs. Theor. Appl. Genet. 80: 833-840.

Sorrells, M.E. 1992. Development and application of RFLPs in polyploids. Crop Sci. 32: 1086-1091.

Souza, E. and M.E. Sorrells. 1991a. Relationship among 70 North American oat germplasms: I. Cluster analysis using quantitative characters. Crop Sci. 31: 599-605.

Souza, E. and M.E. Sorrells. 1991b. Relationship among 70 North American oat germplasms: II. Cluster analysis using qualitative characters. Crop Sci. 31: 605-612.

Tivang, J.G., J. Nienhuis, and O.S. Smith. 1994. Estimation of sampling variance of molecular marker data using the bootstrap procedure. Theor. Appl. Genet. 89: 259-264.

Wang, Z.Y., G. Second, and S.D. Tanksley. 1992. Polymorphism and phylogenetic relationships among species in the genus *Oryza* as

determined by analysis of nuclear RFLPs. Theor. Appl. Genet. 83: 565-581.

Weir, B. 1990. Genetic data analysis. Sinauer Associates, Inc. Publishers, Sunderland, MA.

Wright, S. 1978. Evolution and the genetics of populations. IV. Variability within and among natural populations. The University of Chicago Press, Chicago, IL.

Zeven, A. and N.C. Zeven-Hissink. 1976. Genealogies of 14000 wheat varieties. The Netherlands Cereal Centre, Wageningen, CIMMYT, Mexico.

ACKNOWLEDGMENTS

We would like to acknowledge Cargill Seeds Incorporated for partial support for this project and USDA Plant Genome National Research Initiative for subcontract #92-GO161-Cornell of USDA NRI Grant No.92-37300-7550. We thank Dr. Charles McCulloch for valuable suggestions on the statistical methodology.

Table 1: Half-size of a (1-α)% CI for ϕ required to obtain a (1-α)% CI for DI with half-size E.

		E		
DI	f	0.10	0.05	0.01
0.10	0.01	0.0100 *	0.0087	0.0020
0.15	0.02	0.0223	0.0142	0.0030
0.20	0.04	0.0347	0.0194	0.0040
0.25	0.06	0.0458	0.0245	0.0050
0.30	0.09	0.0566	0.0296	0.0060
0.35	0.12	0.0670	0.0346	0.0070
0.40	0.16	0.0774	0.0397	0.0080
0.45	0.20	0.0877	0.0447	0.0090
0.50	0.25	0.0980	0.0497	0.0100
0.55	0.30	0.1082	0.0548	0.0110
0.60	0.36	0.1183	0.0598	0.0120
0.65	0.42	0.1285	0.0648	0.0130
0.70	0.49	0.1386	0.0698	0.0140
0.75	0.56	0.1487	0.0748	0.0150
0.80	0.64	0.1587	0.0798	0.0160
0.85	0.72	0.1688	0.0848	0.0170
0.90	0.81	0.1788	0.0898	0.0180
0.95	0.90	0.1890	0.0948	0.0190

* E = 0.07 for DI = 0.10 because DI is not defined for negative values of f.

Table 2: Sample size for bands to be scored in order to get a 90% confidence interval of size 2E given the genetic distance index (DI) and the number of cultivars to be compared.

		number of cultivars						
DI	E	2	3	5	10	20	30	> 30
0.10	0.07 *	266	441	644	909	1138	1297	1584
	0.05	352	582	851	1201	1503	1714	2093
	0.01	6657	11019	16094	22723	28443	32433	39600
0.20	0.10	86	142	207	293	366	418	510
	0.05	274	454	663	937	1173	1337	1632
	0.01	6455	10685	15606	22034	27581	31451	38400
0.30	0.10	69	114	166	235	294	335	409
	0.05	251	416	608	858	1074	1225	1496
	0.01	6119	10129	14793	20887	26145	29813	36400
0.40	0.10	60	100	146	206	258	294	359
	0.05	229	380	554	783	980	1117	1364
	0.01	5648	9349	13655	19280	24133	27519	33600
0.50	0.10	53	87	127	179	224	256	312
	0.05	204	338	494	697	872	995	1215
	0.01	5043	8348	12192	17214	21548	24571	30000
0.60	0.10	44	73	107	151	189	216	263
	0.05	173	287	419	592	740	844	1031
	0.01	4303	7123	10404	14689	18387	20967	25600
0.70	0.10	35	58	85	119	149	170	208
	0.05	138	228	334	471	589	672	821
	0.01	3429	5676	8291	11706	14652	16708	20400
0.80	0.10	25	41	59	84	105	120	146
	0.05	97	161	235	332	416	474	579
	0.01	2421	4007	5852	8263	10343	11794	14400
0.90	0.10	13	21	31	44	55	63	77
	0.05	51	85	124	175	219	250	305
	0.01	1278	2115	3089	4361	5459	6225	7600

* E = 0.07 for DI = 0.10 because DI is not defined for negative values of f.

Table 3: Distance index (DI), 90% confidence interval size 2E observed, and number of RFLP fragments required to obtain an E = 0.10 in randomly selected subsets of oat.

# cultivars in		E		# fragments	
subsets	DI	pairwise	simultaneous	pairwise	simultaneous
10	0.45	0.04	0.07	56	193
10	0.32	0.04	0.08	67	228
10	0.43	0.04	0.07		198
10	0.43	0.04	0.07	58	198
10	0.43	0.04	0.07	58	198
20	0.43	0.04	0.08	58	248
20	0.43	0.04	0.08	58	248
20	0.42	0.04	0.(	59	251
20	0.40	0.04	0.08	60	258
20	0.42	0.04	0.08	59	251
84	0.40	0.04	0.10	60	359

* Comparisons between 2 cultivars (pairwise) and between 84 cultivars (simultaneous).

Table 4: Characteristics of 34 wheat cultivars surveyed for pedigree and RFLP.

Cultivar	Grain Type	Grain Color	Winter/Spring	Origin
Arthur	Soft	Red	Winter	USA
B17717	Hard	Red	Winter	USA
B617	Hard	Red	Winter	USA
B725S	Hard	Red	Winter	USA
Bezostaya	Hard	Red	Winter	Russia
Blackhull	Hard	Red	Winter	USA
Blueboy	Soft	Red	Winter	USA
Caldwell	Soft	Red	Winter	USA
Centurk 78	Hard	Red	Winter	USA
Era	Hard	Red	Spring	USA
Gaines	Soft	White	Winter	USA
Glenlea	Hard	Red	Spring	Canada
Hart	Soft	Red	Winter	USA
Justin	Hard	Red	Spring	USA
Lac172	Hard	Red	Spring	Argentine
Lac377	Hard	Red	Spring	Argentine
Lac450	Hard	Red	Spring	Argentine
Lac519	Hard	Red	Spring	Argentine
Lac611	Hard	Red	Spring	Argentine
Lac699	Hard	Red	Spring	Argentine
Lac721	Hard	Red	Spring	Argentine
Lac732	Hard	Red	Spring	Argentine
Lac739	Hard	Red	Spring	Argentine
Lancota	Hard	Red	Winter	USA
Lar1138	Hard	Red	Spring	Argentine
Lar193	Hard	Red	Spring	Argentine
Moro	Soft	White	Winter	USA
Olaf	Hard	Red	Spring	USA
Pike	Soft	Red	Winter	USA
R654	Hard	Red	Winter	USA
Selkirk	Hard	Red	Spring	Canada
Sturdy	Hard	Red	Winter	USA
Tascosa	Hard	Red	Winter	USA
Triumph 64	Hard	Red	Winter	USA

Table 5: Distance index (DI), 90% confidence interval size 2E observed, and number of RFLP fragments required to obtain an E = 0.10 in randomly selected subsets of wheat.

# cultivars in		E*		# fragments*	
subsets	DI	pairwise	simultaneous	pairwise	simultaneous
10	0.43	0.05	0.10	58	198
10	0.46	0.05	0.10	56	190
10	0.46	0.05	0.10	56	190
10	0.45	0.05	0.10	56	193
10	0.46	0.05	0.10	56	190
20	0.45	0.05	0.11	56	241
20	0.44	0.05	0.11	57	244
20	0.45	0.05	0.11	56	241
20	0.44	0.05	0.11	57	244
20	0.44	0.05	0.11	57	244
34	0.45	0.05	0.14	56	336

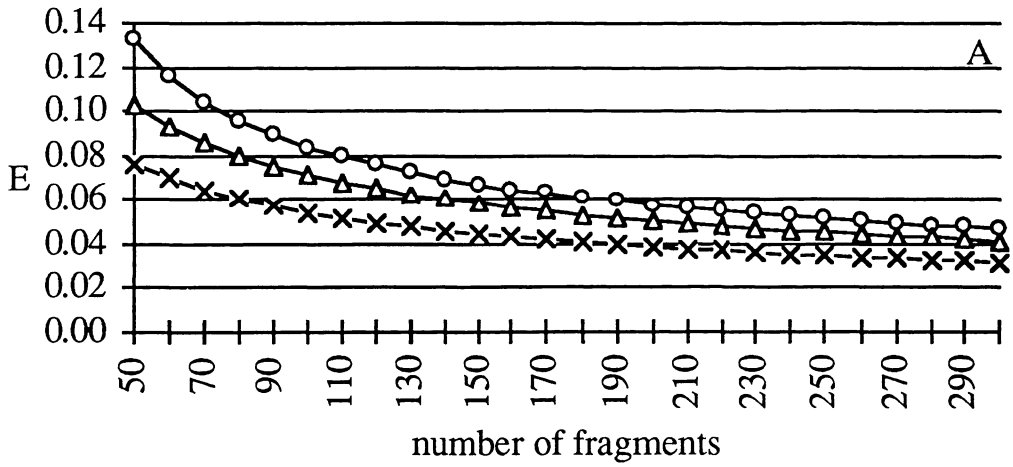
* Comparisons between 2 cultivars (pairwise) and between 34 cultivars (simultaneous).

Fig. 1: Half-size of a 90% CI (E) according to the number of fragments scored and three levels of ϕ (0.25 = O, 0.50 = Δ, and 0.75 = x). A = two cultivars compared; B = > 30 cultivars compared.

Fig. 2: Relationship between the frequency of mismatches (ϕ) with DI (Δ) and Nei's distance (O).

Fig. 3: Dendogram of 34 wheat cultivars compared using RFLP based distance DI.

Fig. 4: Dendogram of 31 wheat cultivars compared using coefficient of parentage.


A
E
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
50
70
90
110
130
150
170
190
210
230
250
270
290
number of fragments

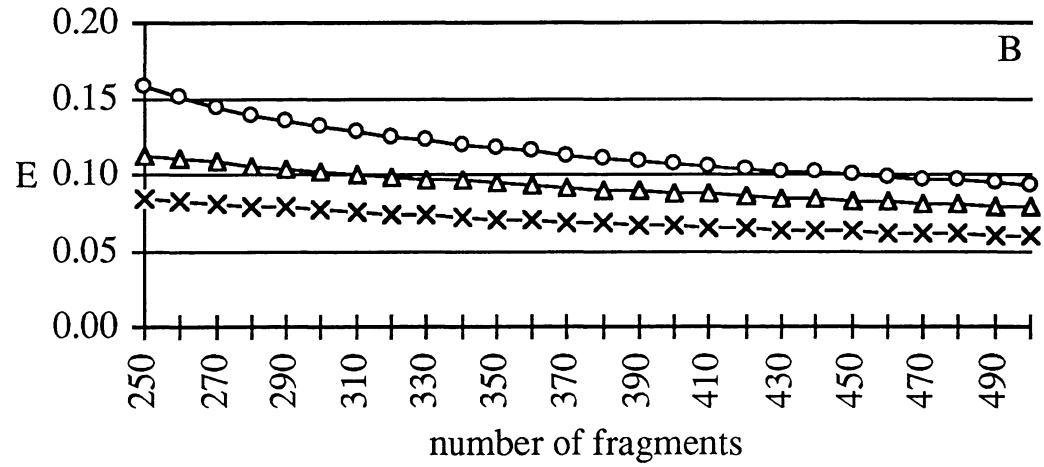

B
E
0.20
0.15
0.10
0.05
0.00
250
270
290
310
330
350
370
390
410
430
450
470
490
number of fragments

Figure 1

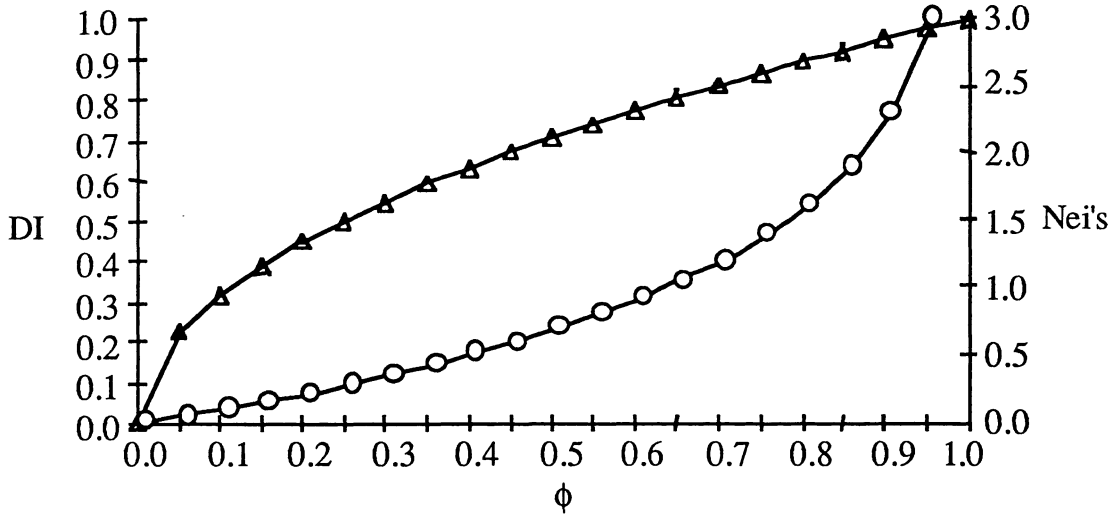

DI
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
ϕ
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Nei's

Figure 2

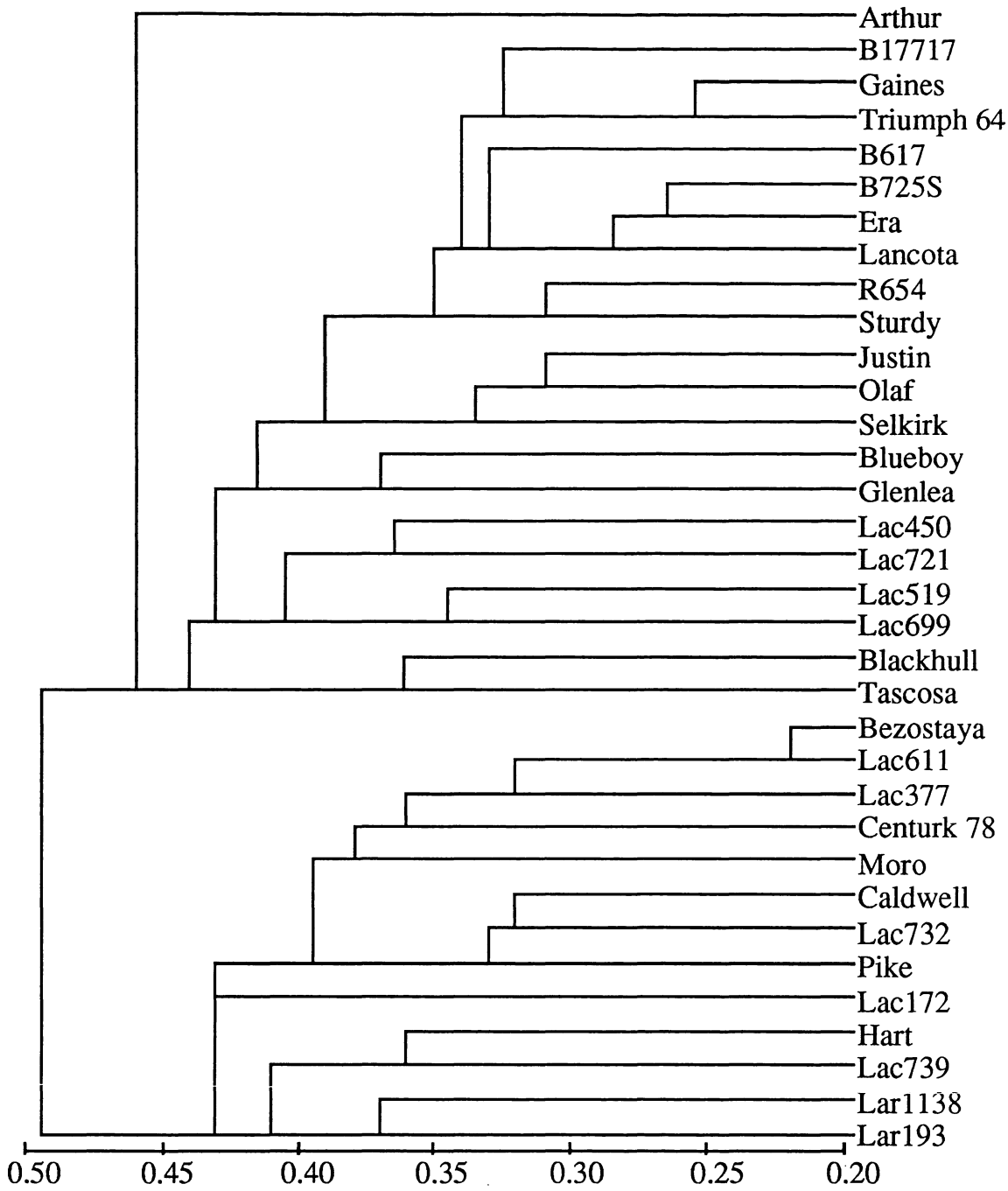

Arthur
B17717
Gaines
Triumph 64
B617
B725S
Era
Lancota
R654
Sturdy
Justin
Olaf
Selkirk
Blueboy
Glenlea
Lac450
Lac721
Lac519
Lac699
Blackhull
Tascosa
Bezostaya
Lac611
Lac377
Centurk 78
Moro
Caldwell
Lac732
Pike
Lac172
Hart
Lac739
Lar1138
Lar193
0.50
0.45
0.40
0.35
0.30
0.25
0.20

Figure 3

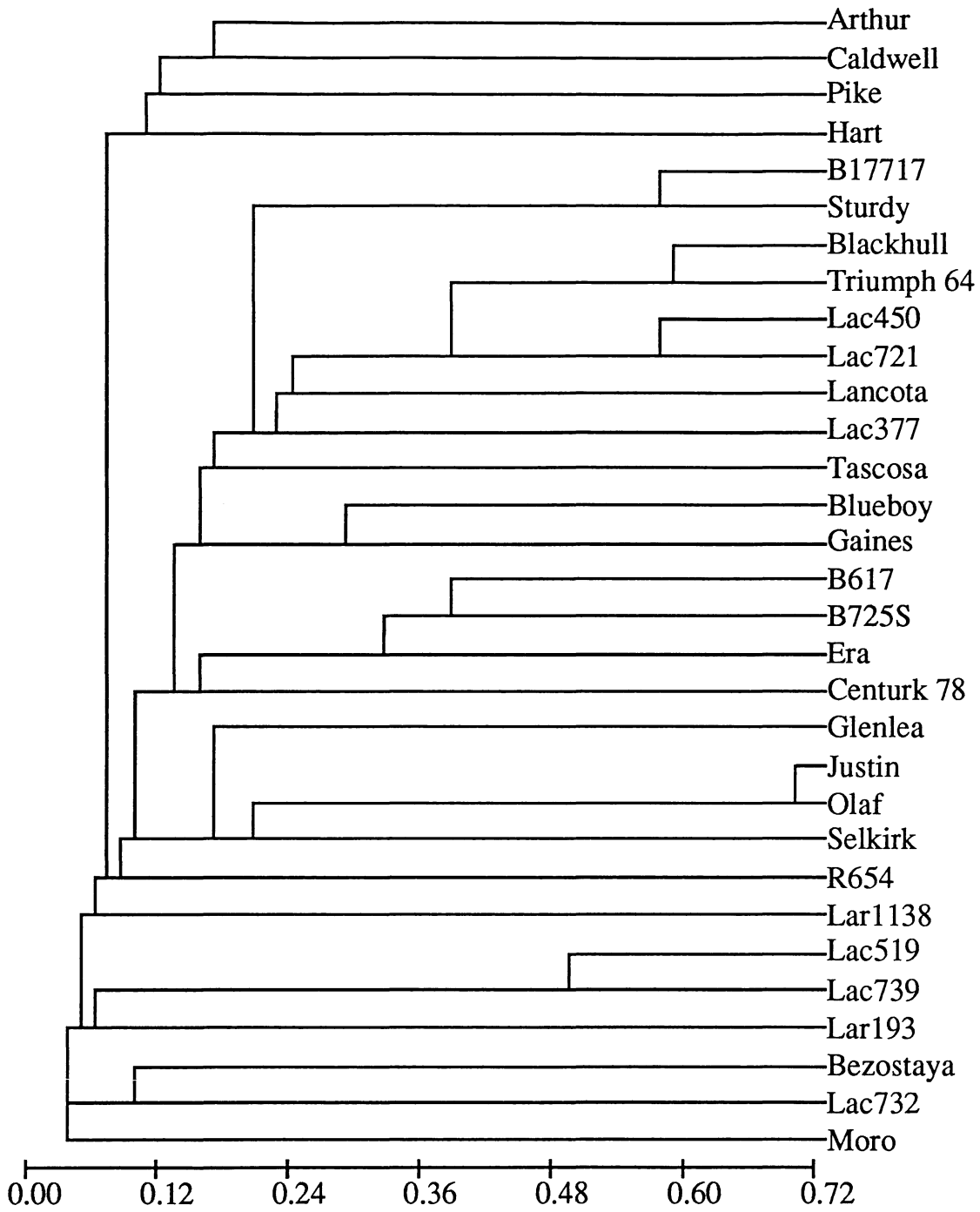

Arthur
Caldwell
Pike
Hart
B17717
Sturdy
Blackhull
Triumph 64
Lac450
Lac721
Lancota
Lac377
Tascosa
Blueboy
Gaines
B617
B725S
Era
Centurk 78
Glenlea
Justin
Olaf
Selkirk
R654
Lar1138
Lac519
Lac739
Lar193
Bezostaya
Lac732
Moro
0.00
0.12
0.24
0.36
0.48
0.60
0.72

Figure 4